Список литературных источников

1. И.В. Бейко. Методы и алгоритмы решения задач оптимизации. Бейко И.В., Бублик Б.Н., Зинько П.Н. – К.: Вища школа. Головное издательство, 1983 – 512 с.
2. Интернет-учебник по курсу «Методы оптимизации» http://kek.ksu.ru/EOS/MO/uchebnik.asp
3. Статья «Программирование и научные вычисления на языке Python». http://ru.wikiversity.org/wiki/
4. Документация SymPy http://sympy.org/en/index.html

ЗАКОН БЕНФОРДА И АТРИБУЦИЯ ТЕКСТОВ

УДК: 51-78, 519.234.3, 519.257, 81-139

Зенков А. В. к.ф.-м.н.доцент
кафедра моделирования управляемых систем
Уральский федеральный университет, ВШЭМ
Казанцев М.В. студент
кафедра моделирования управляемых систем
Уральский федеральный университет, ВШЭМ

Аннотация. Исследовано распределение первой значащей цифры в числительных связных текстов. Обнаружено, что закон Бенфорда приближённо выполняется для них. Отклонения от закона Бенфорда являются статистически устойчивыми авторскими особенностями, позволяющими при некоторых условиях различить части текста с разным авторством.

Ключевые слова: закон Бенфорда, статистическая проверка гипотез, критерий Манна-Уитни.

Abstract. The distribution of the first significant digit in numerals of connected texts is considered. Benford's law is found to hold approximately for them. Deviations from Benford's law are statistically significant author peculiarities that allow, under certain conditions, to distinguish between parts of the text with a different authorship.

Keywords: Benford’s law, Statistical hypothesis testing, Mann – Whitney U-test.

Введение

Своеобразное проявление Закона больших чисел – известный уже более ста лет закон Бенфорда [1] – в последние десятилетия из статистического курьёза превращается в полезное средство анализа данных. Этот закон описывает вероятность появления определённой первой значащей цифры в разнообразных распределениях величин, взятых из реальной жизни. Вопреки здравому предположению о том, что частоты появления любой первой значащей цифры должны быть равными, для многих массивов данных в качестве первой значащей цифры чаще других встречается единица! Согласно закону Бенфорда вероятность появления цифры d в качестве первой значащей

$$P\ d\ = \lg\ 1+\frac{1}{d}\ , \qquad (1)$$

так что $d=1$ должна встречаться с вероятностью $\lg 2 \approx 0{,}30$, $d=2$ – с вероятностью 0,18 и т.д.

Исчерпывающего объяснения закона Бенфорда, охватывающего все случаи реализации, до сих пор не предложено, хотя и сформулированы некоторые условия, благоприятствующие его появлению. Один из классических опытов Бенфорда, хорошо согласующийся с (1) – анализ встречаемости числительных в статьях случайно выбранного выпуска популярного журнала – находит логичное объяснение в теореме Хилла [2], согласно которой в условиях неоднократного случайного выбора распределения вероятностей с последующим случайным выбором числа согласно этому распределению возникает набор чисел, подчиняющийся закону Бенфорда. Как мы покажем ниже, эти условия не являются необходимыми: даже для связного текста, к которому условия теоремы неприменимы, наблюдается распределение первых значащих цифр числительных, близкое к (1).

Несмотря на неполноту объяснения закон Бенфорда успешно применяется для выявления подлогов в бухгалтерской отчётности [3] и фальсификаций на выборах [4]; обсуждаются применения в различных областях от сейсмологии [5] до стеганографии [6]. Одним из авторов настоящей работы показана перспективность использования закона Бенфорда для задач текстологии [7].

Цель настоящей работы – показать, что при определённых условиях исследование частот появления различных первых значащих цифр в связном тексте может быть полезным с точки зрения вопроса об авторстве текста, поскольку эти частоты специфичны для автора.

Статистическое исследование текстов

Для поставленных целей наиболее показателен анализ разных произведений одного автора либо произведений разных авторов на близкие темы.

Исследованию были подвергнуты:

1) записки Юлия Цезаря о Галльской войне и о Гражданской войне;
2) четыре канонических Евангелия – от Матфея, Марка, Луки и Иоанна.

Для этих текстов исследовались частоты появления различных первых значащих цифр с учётом количественных и порядковых числительных, выраженных как цифрами, так и (значительно чаще) словесно.

Известно, что первые семь книг Записок о Галльской войне написаны Цезарем, а последняя, восьмая книга – Гирцием, завершившим произведение. На рис. 1–4 приведены распределения частот встречаемости первой значащей цифры, характерные для Цезаря и Гирция. При общем приближённом выполнении закона Бенфорда (1) легко заметны характерные авторские различия. Всему произведению свойственно более редкое появление единицы в качестве первой значащей цифры по сравнению с предписанием закона Бенфорда и обратное явление для цифры 2 (Рис. 1). Это авторская особенность стиля Цезаря: она проявляется в принадлежащих его перу книгах Записок (на Рис. 2, 3 представлены для сравнения результаты для первой и третьей книг

Записок). Совершенно иной вид имеет аналогичное распределение для восьмой книги Записок, автором которой является Гирций (Рис. 4).

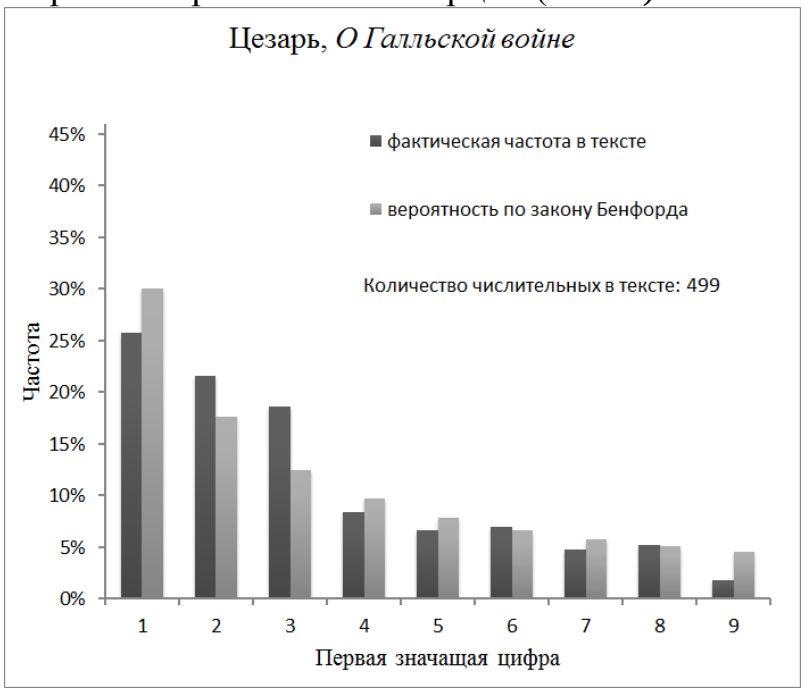

Рис. 1. Распределение первых значащих цифр числительных в Записках о Галльской войне Цезаря. Результаты обработки всех 8 книг Записок

В другом произведении Цезаря – Записках о Гражданской войне, написанных им единолично, – наблюдается та же характерная особенность (Рис. 5). Поскольку это произведение посвящено иным историческим событиям, данное совпадение следует расценивать как отражение именно *стиля* Цезаря.

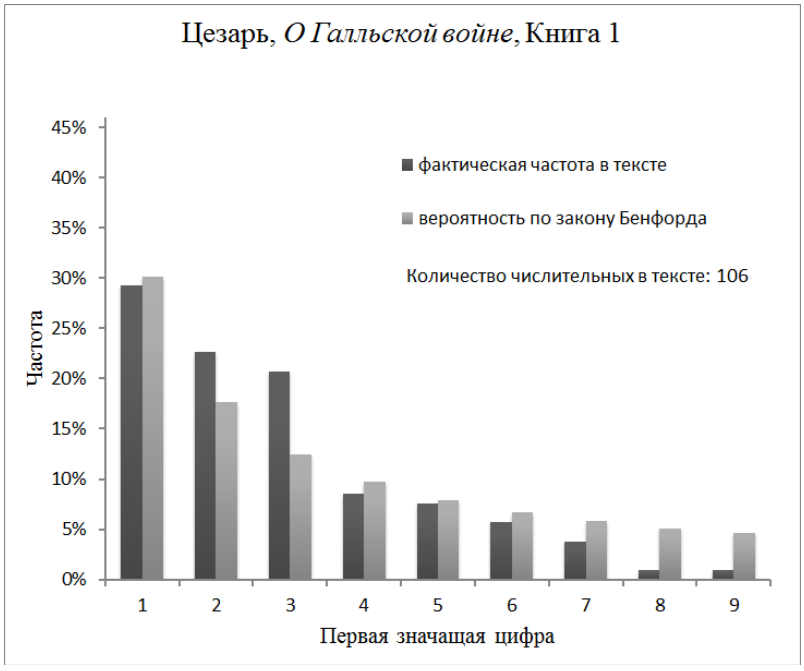

Рис. 2. Распределение первых значащих цифр числительных в Записках о Галльской войне Цезаря. Результаты обработки только 1-й книги Записок, написанной Цезарем

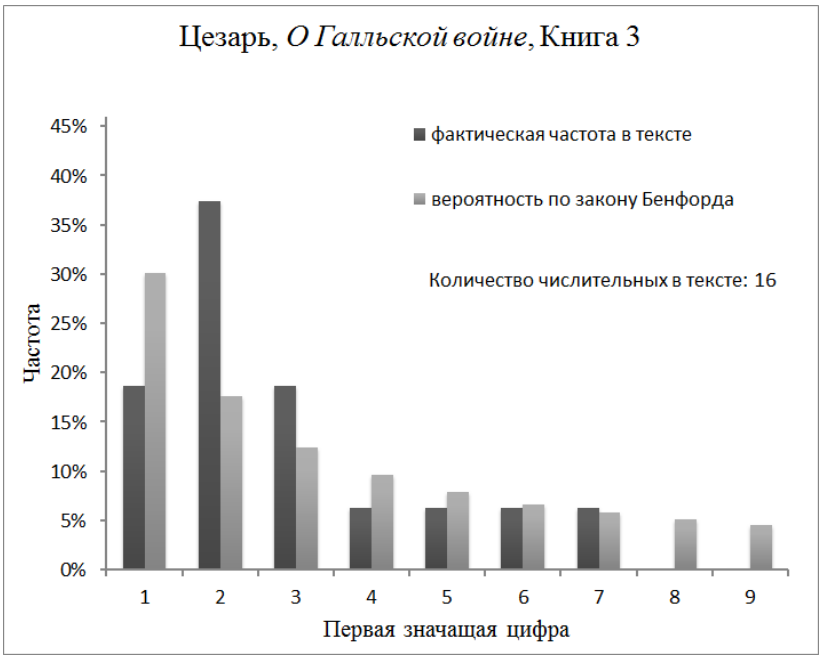

Рис. 3. Распределение первых значащих цифр числительных в Записках о Галльской войне Цезаря. Результаты обработки только 3-й книги Записок, написанной Цезарем

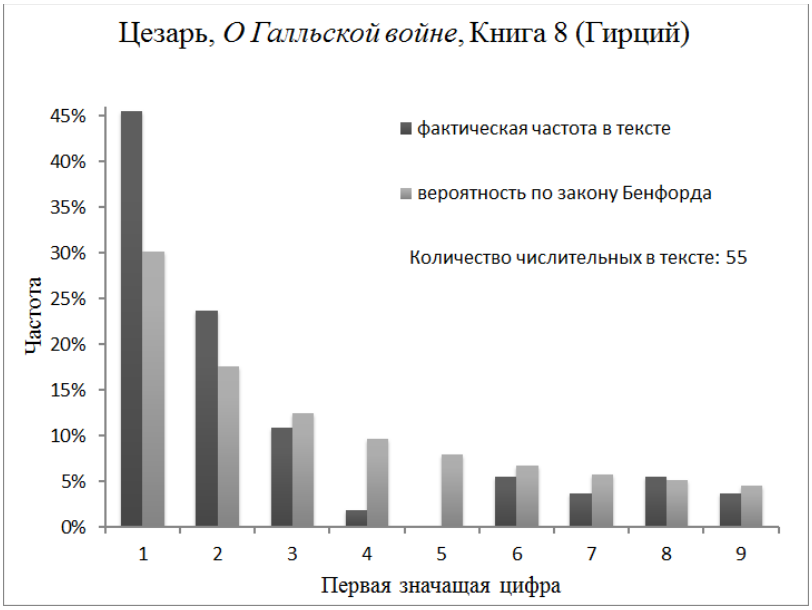

Рис. 4. Распределение первых значащих цифр числительных в Записках о Галльской войне Цезаря. Результаты обработки только 8-й книги Записок, написанной Гирцием

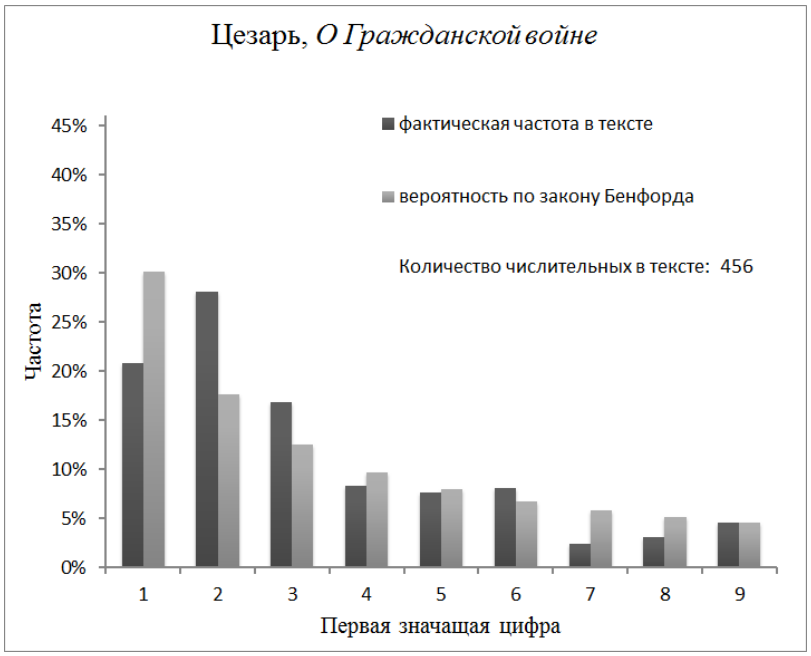

Рис. 5. Распределение первых значащих цифр числительных в Записках о Гражданской войне Цезаря

В качестве ещё одного примера произведений разных авторов с общей тематикой нами рассмотрены Евангелия от Матфея, Марка, Луки и Иоанна. В целом распределение первой значащей цифры числительных и здесь напоминает распределение Бенфорда, но заметно преобладает единица (рис. 6–9). Между распределениями (особенно первыми тремя и последним) наблюдаются различия – не очень большие, но статистически значимые, с учётом объёма проанализированных данных.

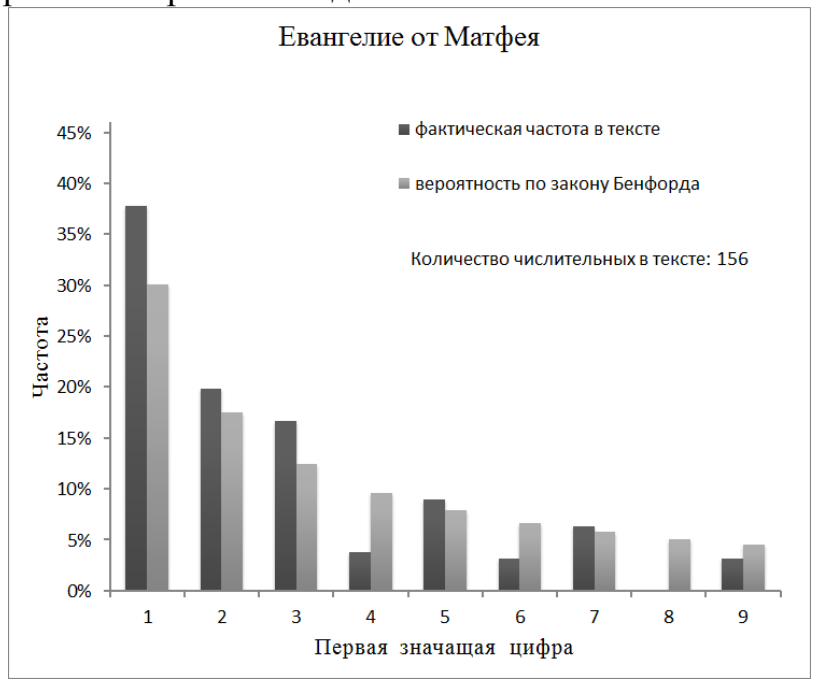

Рис. 6. Распределение первых значащих цифр числительных в Евангелии от Матфея

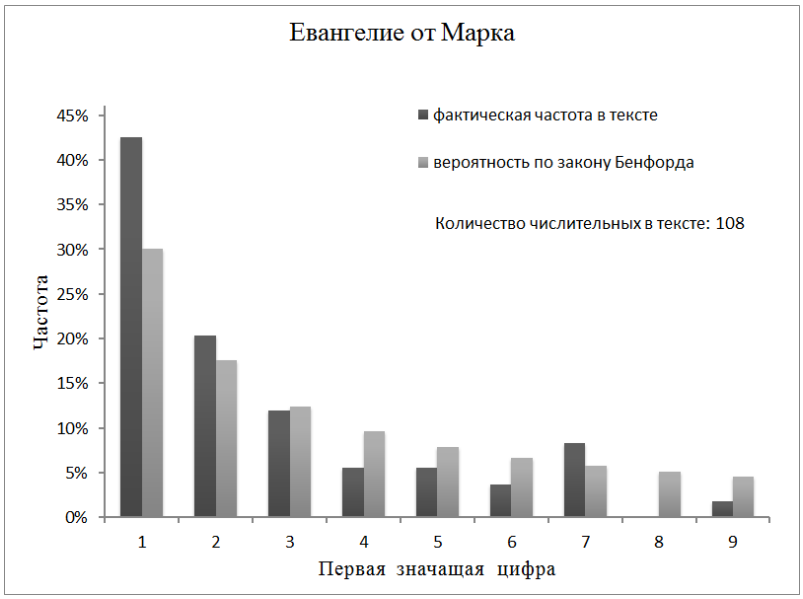

Рис. 7. Распределение первых значащих цифр числительных в Евангелии от Марка

Известное нам исследование [8] распределения первых значащих цифр числительных в корпусе Священного Писания преследовало иные цели.

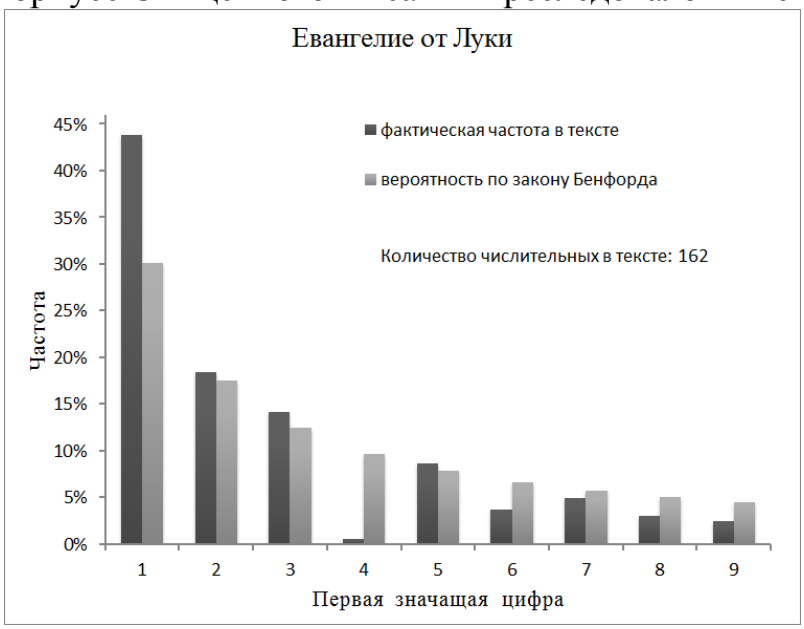

Рис. 8. Распределение первых значащих цифр числительных в Евангелии от Луки

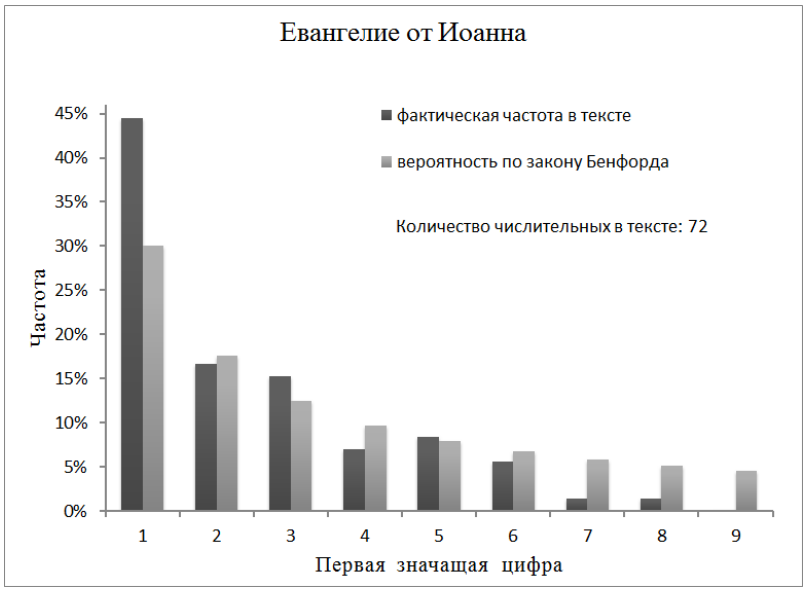

Рис. 9. Распределение первых значащих цифр числительных в Евангелии от Иоанна

Разумеется, сравнение распределений не может основываться только на выявлении субъективных визуальных сходства и различий между ними. Нами применён непараметрический U-критерий Манна-Уитни. Нулевая гипотеза H_0, утверждающая *отсутствие* значимых различий в рассмотренных распределениях, оказалась отвергнутой и принятой именно в тех случаях, как описано выше. А именно, различие между книгами записок о Галльской войне, написанными Цезарем и Гирцием, оказалось значимым ($p = 0,02$), а между разными книгами Цезаря – не значимым ($p = 0,25$). Различия между Евангелиями от Матфея, Марка, Луки, с одной стороны, и Евангелием от Иоанна, с другой, оказались значимыми ($p = 0,03 \div 0,04$), а между любыми двумя Евангелиями из первой тройки – не значимыми. Итак, предлагаемый нами метод разграничения авторства не всесилен, но может быть полезным дополнением к традиционным методам [9].

Заключение

Закон Бенфорда приближённо выполняется для связных текстов.

Отклонения от закона Бенфорда являются статистически значимыми авторскими особенностями, позволяющими при некоторых условиях различить части текста с разным авторством. Очевидными требованиями является достаточная длина текста и употребительность числительных в нём, чему, например, как правило, удовлетворяет историческая литература.

Распределение цифр конца ряда $1,2,\ldots,7,8,9$ подвержено сильным флуктуациям и непоказательно.

Список литературных источников

[1] Benford F. The law of anomalous numbers // Proceedings of American Philosophical Society. – 1938. – vol. 78, No. 4. – P. 551–572.

[2] Hill T.P. A Statistical Derivation of the Significant-Digit Law // Statistical Science. – 1995. – vol. 10 – P. 354–363.

[3] Nigrini M.J. Benford’s Law: applications for forensic accounting, auditing, and fraud detection. – Hoboken: John Wiley & Sons, Inc., 2012. – XX + 330 pp.

[4] Battersby S. Statistics hint at fraud in Iranian election // New Scientist. – 24 June 2009.

[5] Sambridge M., Tkalčić H., Arroucau P. Benford’s Law of First Digits: from Mathematical Curiosity to Change Detector // Asia Pacific Mathematics Newsletter. – 2011. – vol. 1, No. 4. – P. 1–6.

[6] Andriotis P., Oikonomou G., Tryfonas T. JPEG steganography detection with Benford’s Law // Digital Investigation. – 2013. – vol. 9, No. 3–4. – P. 246–257.

[7] Зенков А.В. Отклонения от закона Бенфорда и распознавание авторских особенностей в текстах // Компьютерные исследования и моделирование. – 2015. – Т. 7, вып. 1. – С. 197–201.

[8] Hüngerbühler N. Benfords Gesetz über führende Ziffern: wie die Mathematik Steuersündern das Fürchten lehrt // EducETH, Publikation der Eidgenössischen Technischen Hochschule Zürich. – 2007. – www.educ.ethz.ch/unt/um/mathe/ana/benford

[9] Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge (Mass.) – London: The MIT Press, 1999. – XXXVII + 680 p.

ИСПОЛЬЗОВАНИЕ ВИРТУАЛИЗАЦИИ В УЧЕБНОМ ПРОЦЕССЕ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ОБУЧЕНИЯ

УДК 004.053

Д.Г. Ермаков к. ф.-м. н., доцент
кафедра анализа систем и принятия решений
Уральский федеральный университет, ВШЭМ
А.В. Присяжный к.т.н., доцент
кафедра анализа систем и принятия решений
Уральский федеральный университет, ВШЭМ
О.Е. Хорев студент
кафедра анализа систем и принятия решений
Уральский федеральный университет, ВШЭМ

Аннотация: данная публикация посвящается решению проблемы повышения наглядности преподаваемых предметов в условиях ограниченных ресурсов или их экономии. Еще один взгляд на виртуальные машины в процессе обучения и необычное использование торрентов.

Ключевые слова: Виртуализация, VirtualPC, VmWare, VirtualBox.

Experience and perspectives of virtualization technologies in the learning process of students. Current problems and decision ways.

Key words: Virtualization, VirtualPC, VmWare, VirtualBox.